Probabilistic Methods for Link-Based Classification at INEX 2008

Luis M. de Campos, Juan M. Fernández-Luna,
Juan F. Huete, and Alfonso E. Romero

Departamento de Ciencias de la Computación e Inteligencia Artificial
E.T.S.I. Informática y de Telecomunicación, Universidad de Granada,
18071 – Granada, Spain
{lci,jmfluna,jhg,aeromero}@decsai.ugr.es

Abstract. In this paper we propose a new method for link-based classification using Bayesian networks. It can be used in combination with any content only probabilistic classsifier, so it can be useful in combination with several different classifiers. We also report the results obtained of its application to the XML Document Mining Track of INEX'08.

1 Introduction

The 2008 edition of the INEX Workshop was the second year that members of the research group "Uncertainty Treatment in Artificial Intelligence" at the University of Granada participate in the Document Mining track. Our aim is as in previous editions, to provide a solution to the proposed problems on the framework of Probabilistic Graphical Models (PGMs).

The corpus given for 2008 differs slightly on the one of the previous year [3]. Again, as in 2007, it is a single-label corpus (a subset of the AdHoc one [2] but using a different set of 16 categories). Moreover, this year a file with the list of links between XML documents has been added. Because the track has been substantially changed, it would be firstly interesting to check the utility of using the link information. We will show later that those links add relevant information for the categorization of documents.

On the other hand, given that the 2008 corpus is coming from the same source (Wikipedia) as the 2007 corpus, we think that it might not be worthwhile to use the structural information of the documents for categorization. In [1] we showed that even using some very intuitive XML document transformations to flat text documents, classification accuracy was not improving, getting worse in some of the cases. In this year, then, we have used a more pragmatic approach, directly ignoring the structural information by simply removing XML tags from the documents.

This work is structured as follows: firstly we perform a study of the link structure in the corpus, in order to show its importance for categorization. Secondly, we present our model for categorization, based on Bayesian networks, with some variants. Then, we make some experiments on the corpus to show how our model performs, and finally, we list some conclusions and future works.

S. Geva, J. Kamps, and A. Trotman (Eds.): INEX 2008, LNCS 5631, pp. 453–459, 2009.

2 Linked Files. Study on the Corpus

As we said before, we are given a set of links between document files as additional training information, making some explicit dependencies arise between documents. This information violates a “natural” assumption of traditional classification methods: the documents are independent of each other. This case of non independent documents can have different forms; and the graph of relationships among documents are not neccesarily regular (it can be a general directed graph, neither a tree nor a forest).

Clearly, for the problem of document categorization, these intra-corpus dependences could be ignored, applying “traditional” text categorization algorithms but, as we will show afterwards, the information from linked files can be a very valuable data.

But, how are those links supposed to help in the final process of text categorization? Obviously, not all kinds of links are equal, because they can give different information (even none). A careful review of those different kinds of dependencies represented by hyperlinks (*regularities*) is given by Yang [6], and following her terminology we can conjecture that we are in a “encyclopedia regularity”. We reproduce here her definition:

> *One of the simplest regularities is that certain documents with a class label only link with documents with the same class label. This regularity can be approximately found in encyclopedia corpus, since encyclopedia articles generally reference other articles which are topically similar.*

We have plotted, in figure 2, a matrix where the rows and columns are one of the 16 categories. Each matrix value $m_{i,j}$ represents the probability that a document of class i links a document of class j, estimated from the training document collection.

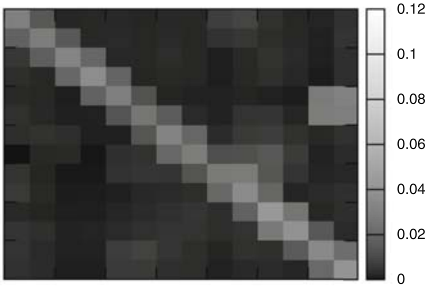

Fig. 1. Probability that a document from category i links a document from category j

As it can be seen (the matrix has a strong weight in its diagonal), documents of one category tend to link documents of the same category. Moreover, doing the same plot with the probability of a document of class i being linked by a document of class j, and another one with the probability of a document of class i links or is linked by a document of class j, we obtain a similar result (a matrix with a high weight for the diagonal values).

Thus, although we could think that only the outlinks tend to be useful, we can affirm that also inlinks are useful, and also consider the links without any direction.

3 Proposed Method

3.1 Original Method

The method proposed is an extension of a probabilistic classifier (we shall use in the experiments the Naive Bayes classifier but other probabilistic classifiers could also be employed) where the evidence is not only the document to classify, but this document together with the set of related documents. Note that, in principle, we will try to use only information which is available in a natural way for a text classifier. Considering that different documents are processed through batch processing, the information easily available to a system, given a document, is the set of documents it links (not the set of documents that link it).

Consider a document d_0 which links with documents $d_1, \ldots, d_m$. We shall consider the random variables $C_0, C_1, \ldots, C_m$, all of them taking values in the set of possible category labels. Each variable C_i represents the event "The class of document d_i is". Let e_i be the evidence available concerning the possible classification of each document d_i (the set of terms used to index the document d_i or the class label of d_i). The proposed model can be graphically represented as the Bayesian network displayed in figure 2.

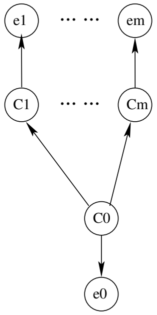

Fig. 2. Bayesian network representing the proposed model

The independencies represented by the Bayesian network are the following: given the true class of the document we want to classify, the categories of the linked documents are independent among each other. Moreover, given the true category of a linked document, the evidence about this category due to the document content is independent of the original category of the document we want to classify.

Our objective is to compute the posterior probability $p(C_0|e)$, where e is all the available evidence concerning document d_0, $e = \{e_0, e_1, \ldots, e_m\}$. It can be proven that, taking into account the independencies represented in the Bayesian network, this probability can be expressed as follows:

$$p(C_0 = c_0|e) \propto p(C_0 = c_0|e_0) \prod_{i=1}^{m} \left(\sum_{c_i} p(C_i = c_i|c_0) \frac{p(C_i = c_i|e_i)}{p(C_i = c_i)} \right) \tag{1}$$

As we can observe in equation (1), the posterior probability of C_0 has two components: a part which only depends on the evidence associated to the document d_0 to be classified ($p(C_0 = c_0|e_0)$) and another part related with the information about the class labels of each one of the documents linked with d_0, which can be obtained using its own local evidence ($p(C_i = c_i|e_i)$). This information is combined with the estimated probabilities of a linked document being of class c_i given that the document linking to it is of class c_0.

The posterior probabilities $p(C_0 = c_0|e_0)$ and $p(C_i = c_i|e_i)$ can be obtained using some standard probabilistic classifier, whereas the probabilities $p(C_i = c_i)$ and $p(C_i = c_i|c_0)$ can be estimated from the training data simply by following these formulas:

$$p(C_i = c_i) = \frac{N_i}{N}$$

and

$$p(C_i = c_i|c_0) = \frac{L_{0i} + 1}{L_{0\bullet} + |C|}$$

where N_i is the number of training documents classified by category i, N is the total number of documents, L_{0i} is the number of links from documents of category 0 to category i, $L_{0\bullet}$ is the total number of links from documents of category 0, and $|C|$ is the number of categories. Note that in the estimation of $p(C_i = c_i|c_0)$ we have used Lapace smoothing. In all our posterior experiments, using Laplace gives better results than not using it.

Therefore, we can think of the proposed model as a method to modify the results offered by a base probabilistic classifier taking into account the information available about the linked documents and the relationships between categories (the prior probabilities $p(C_i = c_i)$ and the values $p(C_i = c_i|c_j)$).

3.2 Extension to Inlinks and Undirected Links

The independencies represented by the Bayesian network are not directly related with the direction of the links. Instead of outlinks, we could think of the previous model as a model that takes into consideration the incoming links. Thus, the

e_i $(i > 0)$ variables would represent the documents that link to one (instead of the files linked by one), and the formula (1) would still be valid. In the case of the incoming links, we should reestimate the dependencies among categories as follows:

$$p(C_i = c_i|c_0) = \frac{L_{i0} + 1}{L_{\bullet 0} + |C|}$$

where L_{i0} is, as previously stated, the number of links from documents of category i to category 0, and $L_{\bullet 0}$ is the total number of links to documents of category 0.

Moreover, in the collective classification literature, the direction of the links is often not considered, so, we also could propose a model where e_i $(i > 0)$ represent the documents linked or being linked (that is to say, neighboured) by the document to classify. In that case, the probabilities would be these:

$$p(C_i = c_i|c_0) = \frac{L_{i0} + L_{0i} + 1}{L_{\bullet 0} + L_{0\bullet} + |C|}.$$

Therefore, these would be our three models: the original one (with incoming links), and the extensions using outlinks and undirected links.

4 Experimental Results

To make the values comparable with the submitted runs, we have also performed some experiments on the test set in order to show the effectiveness (recall) of our approach. First of all we study the two submitted runs, a baseline (flat text classifier) and our proposals (combined with Naive Bayes):

- A classical Naive Bayes algorithm on the flat text documents: 0.67674 of recall.
- Our model (outlinks): 0.6787 of recall.

The two aditional models which were not submitted to the track give the following results:

- Our model (inlinks): 0.67894 of recall.
- Our model (neighbours): 0.68273 of recall.

Although all our methods improve the baseline, the results achieved are not really significant. In order to justify the value of our model, we are asking now ourselves which is the predicting power of our proposal, by making some additional computations in an "ideal setting". This "ideal setting" is, for a document being classified, to be surrounded (linking, linked by or both of them) with documents whose class membership is perfectly known (and hence we can set for a related document d_k of category c_i, $P(C_k = c_i|d_k) = 1$ -the true class- and $P(C_k = c_j|d_k) = 0$ -the false categories- $\forall c_j \neq c_i$). Remember that, in previous experiments, a surrounding file whose category was not known should be first classified by Naïve Bayes, and then that estimation (the output probability values) was used in our model.

So, the procedure is the following: for each document to classify, look at the surrounding files. For each one, if it is a training file, use that information (perfect knowledge), and if it is a test file, use also its categorization information taken from the test set labels to have our file related to documents with perfect knowledge. This "acquired" knowledge is obviously removed for the next document classification.

In this "ideal setting" we have made two experiments: one combining naïve Bayes with our model (like the second one of the previous two), and one which combined a "blind classifier" (the one that gives equal probability to each category) with our model. The first should be better than the two previous ones, and the second one could give us an idea of the true contribution to the predictive power of our model, despite the underlying basic classifier used.

– Model for outlinks in an "ideal setting" using Naive Bayes as a base classifier: 0.69553 of recall.
– Model for outlinks in an "ideal setting" using a "blind classifier": 0.46500 of recall.
– Model for inlinks in an "ideal setting" using Naive Bayes as a base classifier: 0.69362 of recall.
– Model for inlinks in an "ideal setting" using a "blind classifier": 0.73278 of recall.
– Model for neighbours in an "ideal setting" using Naive Bayes as a base classifier: 0.70212 of recall.
– Model for neighbours in an "ideal setting" using a "blind classifier": 0.66271 of recall.

The first experiment provides the desired result: the recall is improved (although not so much). The small improvement could be due, in some part, to the extreme values given in this corpus by the Naive Bayes classifier (very close to 0 and 1). The introduction of these values in the final formula, as the first factor in the final posterior probability of each document, makes difficult to take into account (in the categories of the values close to 0) the information provided by the second factor (the combination of the information given by all the linked files), vanishing in some cases because of the low value of the first factor.

However, the second experiment showed us that, only using link information, and ignoring all content information of the document to classify, in this "ideal setting" of knowing the true class of each surrounding document, our method can reach 0.46500, 0.73278 or 0.66271 of recall. In the case of the inlinks, ignoring the content of the document to classify and perfectly knowing the values of the categories of the surrounding documents, gives better results than using this content. Besides, these values are clearly high, whichg gives us the idea of the predictive power of link information in this problem.

5 Conclusions and Future Works

We have proposed a new model for classification of linked documents, based on Bayesian networks. We have also justified the possibly good performance of the model in an "ideal" environment, with some promising results. Regrettably, our results in this track have been very discrete, reaching the final positions and not improving so much the naïve Bayes baseline.

To improve those poor results in the future, we could use a classifier (probabilistic) with a better performance. Such a classifier could be a logistic regression procedure, a higher dependence network or just a SVM with probabilistic output (using Platt's algorithm [5]). The probability assignments should also be "softer", in the sense that several categories should receive positive probability (naïve Bayes tended to concentrate all the probability in one category, zeroing the others and making the information provided by the links not useful, in some way).

As future work we would like to study this problem as a collaborative classification problem (see, for instance [4], and try to apply this method in one of the particular solutions (those that need a "local classifier") that are being given to it.

Acknowledgments. This work has been jointly supported by the Spanish Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía, Ministerio de Ciencia de Innovación and the research programme Consolider Ingenio 2010, under projects TIC-276, TIN2008-06566-C04-01 and CSD2007-00018, respectively.

References

1. de Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Romero, A.E.: Probabilistic Methods for Structured Document Classification at INEX 2007. In: Fuhr, N., Kamps, J., Lalmas, M., Trotman, A. (eds.) INEX 2007. LNCS, vol. 4862, pp. 195–206. Springer, Heidelberg (2008)
2. Denoyer, L., Gallinari, P.: The Wikipedia XML Corpus. SIGIR Forum 40(1), 64–69 (2006)
3. Denoyer, L., Gallinari, P.: Report on the XML mining track at INEX 2007 categorization and clustering of XML documents. SIGIR Forum 42(1), 22–28 (2008)
4. Sen, P., Getoor, L.: Link-based Classification, Technical Report CS-TR-4858, University of Maryland, Number CS-TR-4858 - February 2007 (2007)
5. Platt, J.: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: Smola, A., Bartlett, P., Scholkopf, B., Schuurmans, D. (eds.) Advances in Large Margin Classifiers. MIT Press, Cambridge (1999)
6. Yang, Y., Slattery, S.: A study of approaches to hypertext categorization. Journal of Intelligent Information Systems 18, 219–241 (2002)